CLAIMS

We claim:

1. A method of providing a scalable multicast infrastructure for multicast messaging on an overlay network including a set of nodes, wherein each node in the set of nodes has a node name indicating a network region of the node, the method comprising:

disseminating messages through a multicast tree formed from a subset of the set of overlay nodes, wherein a root node of the multicast tree belongs to a first network region and a path in the multicast tree is prohibited from re-entering the first network region once the path leaves the first network region.

2. The method of claim 1, wherein the multicast tree is formed by routing a subscription message from a subscriber node in the first network region to the root node, comprising:

receiving the subscription message at a node in the first network region;

recording a forwarding pointer to a previous node from which the message was received; and

forwarding the message to the root node by routing the message to a next node within the first network region, based on a node name of the next node.

3. The method of claim 1, further comprising:

creating a topic for which messages are published;

forming a plurality of multicast trees;

publishing messages about the topic to a root node of each of the plurality of multicast trees; and

forwarding the messages to subscribers through the plurality of multicast trees.

4. The method of claim 3, wherein a subscriber in the first network region finds the topic using a name service comprising a directory of topics published in the first network region.

5. The method of claim 1, wherein a network region is one of a geographic locality and an administrative domain.

6. The method of claim 1, wherein a network region comprises a subset of the set of overlay nodes, and wherein the network region is owned by an organization and each node in the network region also belongs to the organization.

7. The method of claim 6, wherein the node name comprises an organizational indicator indicating ownership by the organization, and an organization-relative indicator that encodes one of a geographic locality and an administrative subdivision within the organization.

8. The method of claim 6, wherein an external node belonging to a second organization sends a subscription message to the root node of the multicast tree, further comprising:

receiving the subscription message at a last node in the second organization,

recording a forwarding pointer to a previous node from which the message was received at the different node, and

determining that a next hop in a routing path to the root node is to a node not in the second organization; and

modifying the subscription message to request than a node in the first organization forward messages directly to the last node.

9. The method of claim 8, further comprising receiving a confirmation message from the node in the first organization.

10. The method of claim 9, wherein, if no confirmation message is received choosing a different node and forwarding the subscription request to the different node.

11. The method of claim 6, wherein an external node belonging to a second organization sends a subscription message to the root node of the multicast tree by determining an internet protocol address of a node in the first organization using a name service and sending the subscription message from the external node to the node belonging to the first organization using a network transport layer underlying the overlay network.

12. The method of claim 3, further comprising maintaining a buffer at each node of each of the plurality of multicast trees to record recent messages.

13. A method of assigning node names to nodes in a scalable multicast infrastructure for multicast messaging on an overlay network, wherein each node has a node name indicating a network region of the node, the node name include an organizational indicator and an organization-relative indicator, the method comprising:

requesting from a global certificate authority a first certificate certifying that an organization owns the organizational indicator;

requesting the global certificate authority to allocate a batch of unique identification numbers to the organization and to issue a second certificate certifying that the organization owns the batch of unique identification numbers;

receiving the first and second certificates;

issuing to a node owned by the organization the first and second certificates;

allocating one of the batch of unique identification numbers to the node;

allocating to the node a node name including the organization indicator and the organization-relative indicator; and

issuing to the node a third certificate, from the organizational certificate authority, certifying that the organization-relative indicator of the node is validly associated with the organization and is bound to the one of the batch of unique identification numbers.

14. The method of claim 13, wherein a network region is one of a geographic locality and an administrative domain.

15. The method of claim 13, wherein a network region comprises a subset of nodes in the overlay network, and wherein the network region is owned by an organization and each node in the network region also belongs to the organization.

16. The method of claim 13, wherein the organization-relative indicator encodes one of a geographic locality and an administrative subdivision within the organization.

17. The method of claim 13, further comprising:

requesting a fourth certificate from the global certificate authority certifying that the organizational certificate authority is authorized to allocate the unique identification number and the node name to the node;

receiving the fourth certificate; and

issuing the fourth certificate to the node.

18. A method of providing a scalable multicast infrastructure for multicast messaging on an overlay network including a set of nodes, the method comprising:

forming a primary multicast tree from a subset of the set of overlay nodes, the multicast tree including a root node and one or more subscriber nodes;

passing an event message from the root node to the one or more subscriber nodes through the primary multicast dissemination tree; and

ensuring delivery of the event message by a first method when message traffic occurs above a predetermined level, and ensuring delivery of the event message by a second method when message traffic is not above the predetermined level.

19. The method of claim 18, wherein the first method includes periodically sending one of an event message and a heartbeat message at a predetermined interval, wherein the one or more subscriber nodes determines a tree disconnect when the one of an event message and a heartbeat message is not received at the predetermined interval.

20. The method of claim 18, wherein the second method includes:

forming a plurality of secondary multicast trees, each secondary multicast tree including a subset of the set of overlay nodes, wherein each secondary multicast tree includes a different root node and shares a common subscriber node that is one of the one or more subscriber nodes of the primary multicast tree;

passing the event message from the root node of the primary multicast dissemination tree and the root node of each secondary multicast dissemination tree to the common subscriber node; and

determining a primary multicast tree disconnect when the event message is received from a root node of one of the plurality of secondary multicast trees and the event message is not received from the root node of the primary multicast tree.

21. The method of claim 20, wherein a full event message is passed from the root node of the primary multicast tree, and a digest of the full event message is passed from the root node of each secondary multicast tree.

22. The method of claim 21, wherein the second method further includes:

sending a request message from the common subscriber node to a root node of one of the plurality of secondary multicast trees when the full event message is not received, the request message indicating that full event messages should be passed to the common subscriber node; and

setting an enable bit at each of a plurality of nodes in a path between the root node of one of the plurality of secondary multicast trees and the common subscriber node, the enable bit indicating that a full event message should be forwarded down the path.

23. The method of claim 20, wherein the root node of the primary multicast tree is located in a same network region as the common subscriber node and the root nodes of the plurality of secondary multicast trees are located at increasing distances from the common subscriber node.

24. The method of claim 23, wherein a network region is one of a geographic locality and an administrative domain.

25. A method of participating in a scalable multicast infrastructure for multicast messaging on an overlay network including a set of nodes, the method comprising:

joining a first multicast tree including overlay nodes in an overlay routing path between a subscriber node and a root node of the first multicast tree; and

joining a second multicast tree formed from the first multicast tree, wherein the second multicast tree includes a subset of the overlay nodes in the first multicast tree, the subset consisting of only nodes that voluntarily participate in message dissemination.

26. The method of claim 25, wherein the first multicast tree includes a plurality of subscribers.

27. The method of claim 25, wherein joining the first multicast tree includes sending a subscription message from a first node addressed to a root node through the overlay network, each node in the overlay routing path:

receiving the subscription message at an intermediate node from a preceding node;

recording a tree forwarding pointer that points to the preceding node at the intermediate node; and

forwarding the subscription message to a next node, wherein the subscription message stops when it reaches one of the root node and another node in the first multicast tree.

28. The method of claim 27, wherein forming the second multicast tree includes assuming forwarding duties of a non-participating node that does not wish to participate in message dissemination, wherein forwarding duties includes forwarding event messages received from a parent node of the non-participating node to a child node of the non-participating node.

29. The method of claim 28, wherein delegating forwarding duties includes:

generating at a participating node a unique delegation ticket that includes a pointer to the participating node; and

passing a unique delegation ticket to an ancestor node of the participating node that is the non-participating node, wherein an ancestor node of the non-participating node is directed to forward messages directly to the participating node.

30. The method of claim 29, further comprising generating only one delegation ticket.

31. The method of claim 29, wherein the non-participating node must pass the delegation ticket to an ancestor node if the ancestor node is also a non-participating node.

32. The method of claim 25, wherein joining the second multicast tree includes:

receiving at the subscriber node a probe message from a node in the second tree, wherein each node in the first tree receiving the subscription message forwards the subscription message through the first tree until the subscription message is received by the node in the second tree; and

sending a message to the node in the second tree instructing the node in the second tree to forward messages directly to the subscriber node.

33. The method of claim 25, wherein, when the subscription message is received at a first node in the first tree, the first node forwards the subscription message to a parent node of the first node if the first node is not a node in the second tree and has not previously forwarded a subscription message to the parent node.

34. The method of claim 25, wherein, when the subscription message is received at a first node in the first tree, the first node forwards the subscription message to a child node of the first node if the first node not a node in the second tree and has previously forwarded a subscription message to the parent node.

35. The method of claim 32, further comprising creating a failure notification group including every node receiving the subscription message, wherein the failure notification group is created using a failure notification service, and wherein the failure notification service removes a relevant state if a failure is ascertained.

36. A computer readable medium having computer-executable instructions for performing the steps of claim 1.

37. A computer readable medium having computer-executable instructions for performing the steps of claim 18.

38. A computer readable medium having computer-executable instructions for performing the steps of claim 25.

5